Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 16:36:41
REFERENCE NO: 291

This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode).

Author Names & Affiliations

- Karen Stocks - Scripps Institution of Oceanography
- Steve Diggs - Scripps Institution of Oceanography
- Matthias Lankhorst - Scripps Institution of Oceanography
- Cheryl Peach - Scripps Institution of Oceanography
- Ilya Zaslavsky - San Diego Supercomputer Center

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Oceanography, information management systems, data science, STEM education and outreach, ocean observatories

Title of Submission

Towards an end-to-end data ecosystem to support oceanographic research, education and outreach

Abstract (maximum ~200 words).

The past 10 years have brought an enormous increase in the accessibility and re-use of oceanographic data, leading to substantial and important research outputs. However, the ability to find and access data of interest remains a challenge. In an environment where in-situ samples are sparse, expensive to collect, and require expert quality control, a concerted effort to create an end-to-end research data ecosystem has the potential to greatly increase the efficiency and productivity of ocean sciences research and education. A central component should be institution-based repositories, where data curators work closely with scientific experts producing the data. This should be supported by the development and application of comprehensive persistent identifier schemes for all first-class research products (publications, data, researchers, awards, code, workflows, ontologies, etc.) to promote provenance tracking and credit. Funding for the creation and curation of key aggregate data products, designed to be accessible to a broad audience, is critical to supporting interdisciplinary and non-expert use. In addition, education and training at a variety of levels, better end-user assessment of existing tools, and research into cyberinfrastructure resource development and sustainability models can improve the rate of CI resource update and persistence in the community.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Data re-use supports efficient and important science, and is particularly important in oceanography, where high-quality data are sparse.

The ocean covers over over 70% of the globe, and is highly undersampled outside of the nearshore zone and the few parameters that can be sensed by satellites; this undersampling limits much oceanographic research. Furthermore, ocean environments are harsh for in-situ instrumentations, where fouling and other stresses limit data quality. With ship time costing tens of thousands of dollars per day for ocean and global class research vessels, and in-situ instrumentation being expensive to deploy and maintain, there is a high incentive to produce the maximum possible benefit from available, high-quality ocean observations. Existing ocean data repositories have shown their value. For example, the Ocean Biogeographic Information System, a global compilation of marine species distribution data, supports 80-100 research publications per year [1].

Furthermore, there are many research challenges that require the integration of large-scale, long-term, or interdisciplinary data, and can only be addressed with data re-use and integration. These include grand challenges such as understanding the processes promoting speciation and diversity in the ocean, and pressing management concerns such as predicting ocean acidification and other climate changes.

Significant advances have been made in the availability, accessibility, and quality of oceanographic data. Programs such as BCO-DMO [2], CCHDO [3], and R2R [4] are working to make research data findable and accessible. NOAA’s NCEI [5] provides centralized access to substantial data resources. Programs like QARTOD [6] are promoting quality control best practices. Organizations like ISO [7] and OGC [8] are developing standards for the description and sharing of data and services.

However, a robust foundation of stable, curated data is still lacking in oceanography, and the geosciences in general. Many data never reach a repository. This limits scientific progress, and contributes to widespread concerns over reproducibility in science [9,10]. EarthCube, which is developing a cyberinfrastructure for geosciences, had a use case working group that interviewed 50 geoscientists to ask what the cyberinfrastructure-related challenges are in their research. 75% of respondents described data access and/or availability as a problem in their personal research - by far the most common concern [11].

Data repositories, even those used and valued by the community, are not always stable. Even widely used, high-profile data resources can be lost or endangered as funding and program priorities change (e.g. CDIAC). This is just one example of a common problem, as budgets and funding priorities fluctuate through time. NSF policy now requires that all products from federally-funded research, including original field data, must be documented and preserved. But new sensors are creating ever-larger volumes of data, and repositories have received largely flat funding, creating a critical pressure that may destabilize more repositories in the future.

Even when data are in a stable repository, there can be duplication of identical or related data across repositories, with insufficient metadata to disambiguate them. For example, the Continuous Plankton Recorder data are available from OBIS, BCO-DMO, and SAHFOS [12]. But each repository supplies a different temporal span, and taxonomic resolution. Unclear or insufficient metadata, and semantic heterogeneity, further complicate the discovery and re-use of data.

Taken together, these challenges create an inefficient and frustrating data landscape for oceanographers, limiting their ability to do integrative and and interdisciplinary research.

REFERENCES

[1] http://iobis.org/library/ Accessed 2017-04-03
[2] Biological and Chemical Oceanography Data Management Office, http://www.bco-dmo.org
[3] CLIVAR and Carbon Hydrographic Data Office, https://cchdo.ucsd.edu
[4] Rolling Deck to Repository, http://www.rvdata.us
[5] National Centers for Environmental Information https://www.ncei.noaa.gov
[6] Quality Assurance of Real Time Oceanographic Data,
https://ioos.noaa.gov/project/qartod/
[7] International Standards Organization, https://www.iso.org
[8] Open Geospatial Consortium, http://www.opengeospatial.org
[9] Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. Nature 533: 452-454. doi:10.1038/533452a
[10] Allison, DB, AW Brown, BJ George, KA Kaiser. 2016. Reproducibility: a tragedy of errors. Nature 530: 27-27. doi:10.1038/530027a
[11] https://www.earthcube.org/group/use-cases-wg. Report available from https://goo.gl/ERhClS. Accessed 2017-04-03
[12] Sir Alistair Hardy Foundation for Ocean Sciences, https://www.sahfos.ac.uk

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

It is critical that NSF foster the development of an end-to-end research data ecosystem with the following key components:

- Sufficient bandwidth to support the collection and transfer of at-sea data. We fully endorse the recommendations of the “University National Oceanographic Laboratory System (UNOLS) Satellite Network Advisory Group (SatNAG) response to NSF CI 2030”.
Active data management and curation by research institutions. We believe that data management is done most effectively by expert data curators with some knowledge of the scientific domain working closely with the scientists who collected and have expertise with the data. On the one hand, current best practices in data management are too complex to expect domain scientists to become proficient: the ISO 19115 metadata standard alone is 150 pages long. It is neither efficient nor reasonable to expect expert oceanographers to learn deep data curation practices. On the other hand, only those experts can fully describe and assess those dataset, and know which are appropriate for different uses. While disciplinary repositories have an important role, it is not feasible to launch a repository for every scientific specialty. If exposed through standard service endpoints, the institutional holdings can be accessed by a diversity of aggregators (NCEI, DataONE, etc.), disciplinary specific portals, and tools/models. We argue that institution-based repositories meeting high community standards (e.g. Data Seal of Approval criteria [13]) are best positioned: institutions are persistent; research programs and PIs are not.

- Comprehensive and machine-crawlable application of persistent identifiers for all first-class research products, including data, researchers (ORCIDs), cruises, publications, awards, code/workflows, and semantic content (ontology classes). This should include the referencing of input data in derived data products. This will solve several problems. First, it will enable credit for scientists producing widely used data products: statistics on the number of times that dataset has been used in other products and, ultimately, in papers, will be available, and can then be used by NSF and institutions to evaluate a researcher’s productivity and confirm that data resulting from an award has been shared. Second, it will enable provenance trees, allowing a researcher to find not just a dataset of interest, but raw and processed versions that may be more appropriate for their use. Third, it will improve the usability of the resources by providing context and semantic integration. This vision will require some research (e.g. PIDs approaches for evolving datasets, which is not yet fully resolved), but primarily the development of agreed practices, supported by researchers, repositories and publishers. COPDESS [14] is making substantial progress in this area.

- Support for the development and curation of key data products. Regardless of how well documented a dataset is, it is not tractable for researchers to use raw data from outside of their domain with confidence. The NOAA World Ocean Database [15] is just one example of a highly-used community product. There is high value in the development and maintenance of quality controlled (see below), aggregate data products, yet funding for this work is difficult to attract. It is also critical that, where feasible, these resources be developed with a low barrier to entry to support broad use in education and outreach as well as research (see Question 3 for further information).

- Support for Data Quality Assessment and Control. Ocean instrumentation, which is deployed in challenging environments and prone to fouling, must be evaluated and quality controlled by experts to be useful to non-experts (or efficiently useful to other experts). This is an unexciting job, but critical to the quality of the science that can be done. It should be a central, supported part of data product generation.

- Research and development of automated methods for data and metadata curation and data quality control. In the foreseeable future, these will not replace the need for supported human experts, but have great potential to advance the state of data management. Expert data curators should be assisted by intelligent systems that will introspect the data, derive additional metadata and describe datasets more rigorously.

REFERENCES
[13] https://www.datasealofapproval.org/en/assessment/ Accessed 2017-04-04
[14] Coalition on Publishing Data in the Earth and Space Sciences. http://www.copdess.org
[15] Boyer, T.P., J. I. Antonov, O. K. Baranova, C. Coleman, H. E. Garcia, A. Grodsky, D. R. Johnson, R. A. Locarnini, A. V. Mishonov, T.D. O'Brien, C.R. Paver, J.R. Reagan, D. Seidov, I. V. Smolyar, and M. M. Zweng. 2013. World Ocean Database 2013, NOAA Atlas NESDIS 72, S. Levitus, Ed., A. Mishonov, Technical Ed.; Silver Spring, MD, 209 pp., http://doi.org/10.7289/V5NZ85MT

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Supporting an end-to-end data management ecosystem will require the following additional aspects:

- Sustained funding. While it is clear that not every CI product can be maintained indefinitely, it is also clear that the current status where valuable resources and tools can disappear because a single person retires, or a program shifts direction, is not sustainable. Some process for identifying and sustaining high-impact, critical resources (including small ones) should be in place.

- Required, external end-user testing and training for all community resources and tools. Many, many cyberinfrastructure components are developed but never taken up by the community. Our experience is that many projects have never put their product in front of an end user who was not part of the project, and would have learned a great deal from doing so. An overemphasis on the computer science/information technology work, and an underemphasis on usability, is both common and financially inefficient. Any resource, whether tool or data repository which is a candidate for funding beyond prototype development should be required to have end user feedback (well beyond a “contact us” link). Furthermore, there should be training available for researchers and students to use CI (e.g. software and/or data carpentry), whether that be NSF hardware, community tools, or useful third party resources.

- Additional research into CI sustainability models, usability, software development models, etc. Given the frankly low success rates of much cyberinfrastructure development, as measured by community uptake, it is critical to fund not just more CI, but research into how to do CI better. What models of sustainability work? What are the lessons learned from project that have successfully developed community products? While there is work in these areas, it should be elevated and expanded.

Pre-college and post-secondary education should not be ignored in creating the cyberinfrastructure for geosciences.

- Integrate data management training into undergraduate and graduate programs. Just as a foundation in statistics is required for almost all advanced science degrees, some foundation in data management should be included in modern degree programs, including quality assessment and control best practices. In concert, there should be further development of courses, degrees, and professional training for multi-faceted data science careers (e.g. data curation, data analytics, etc.).

- Support non-expert access to data repositories and tools. Increasingly, access to ocean sciences and other data related to the earth system is important not only for scientists but for non-expert users such as those with a future in the STEM workforce. Yet expert data interfaces pose a significant barrier to access. Science and math classes at the secondary and post-secondary level are the places where students most often encounter data. Teaching data using skills (“data literacy”) in these classes has become a basic workforce training imperative that requires supporting cyberinfrastructure that provides students with access to the data that constitutes the evidentiary basis for concepts they are studying. As society faces difficult decisions about how to respond to our rapidly changing earth system, a workforce and a populace that understands evidence based geosciences research that is grounded in data and modelling will be better prepared to make decisions in their personal and professional lives. In this context, training and and tools that increase accessibility to geosciences data can influence decisions that impact the future of the planet. Detailed description of the challenges and recommendations for non-expert data access specific to the geosciences can be found in the EarthCube Education End-User Workshop report [16].

REFERENCES
[16] EarthCube Education End-User Report http://nagt.org/nagt/programs/earthcube/index.html. Accessed 2017-04-05

Consent Statement

- “I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode).”